COURSE WORK 1: DATA MINING

MSc Data Science: Coventry University UK (**2024.2 BATCH**)

Name: P D Lolitha Lakshan Weerasinghe

Index Number: COMScDS242P-003

Question 1: Online Retail Data

GitHub Link - Code: <https://github.com/lolitha-lakshan-1/DataMining_assignment_1/blob/main/Assignment.Rmd>

## Task 1 – Data Understanding

As the initial task excel dataset was loaded to R Studio for preliminary analysis. In this step dataset was checked for,

1. How the dataset composed, what are the columns which exists?
   1. It was found all the said columns existed as mentioned which are,

InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

1. Are the column data types being correct?
   1. It was noted that most column types contained correct data types including the “InvoiceDate”. However, it was noted “Stock Code” contained alpha numeric values rather than a 5-digit integral number.
2. What are the missing values?
   1. It was found that “Description” value is missing from 1454 transactions and “CustomerID” is missing from 135080 transactions.
3. What are the unique values in columns such as “Country” and “Stock Code”
   1. “Country” column contained all the valid country values except there were some instances of value being “Unspecified”. There are 446 transactions with country specified as “Unspecified.”
   2. “Stock Code” contained non product codes such as 'BANK CHARGES', 'POST', 'DOT', 'M', 'PADS', 'C2'
4. Are there any outliers in the numeric columns?
   1. “Quantity” column contained values ranging from -80995.00 to 80995.00. However, the mean is 9.55 and median being 3.00. There are 10624 transactions with negative quantities.
   2. “Unit Price” ranges from -11062.06 to 38970.00 while mean being 4.61 and median being 2.08. There are 2 transactions with negative unit prices.

Box Plot of Quantity prior to removing outliers.

A black and white diagram of a number

AI-generated content may be incorrect.

Box Plot of Unit Price prior to removing outliers.

A graph of a number of numbers

AI-generated content may be incorrect.

1. Are there any correlated columns?

A green squares with white text

AI-generated content may be incorrect.

According to the heatmap there were not feature correlations found.

### Data Cleansing

1. Removed transactions which contained negative value for Unit Price or Quantity fields. About 10626 transactions were removed.
2. Removed cancelled transactions which contained “C” in the InvoiceNo.
3. Removed transactions with non-product Stock Codes such as 'BANK CHARGES', 'POST', 'DOT', 'M', 'PADS', 'C2'. About 2324 transactions were removed from the dataset.
4. Removed transactions with Country specified as “Unspecified”. About 446 transactions were removed.
5. It was decided not to remove customer null transactions as it does not affect the basket values.
6. After cleansing there are 528513 transactions remaining on the dataset.

## Task 2 – Apriori Implementation

After the data cleansing dataset was grouped using the combination of Country and the InvoiceNo. This would result unique combinations of Country and InvoiceNo, removing any duplicates. Then it will count the number of distinct InvoiceNo per country, and sort them in descending order of count and would output the top three countries with the number of transactions.

top\_countries <- data %>%

distinct(Country, InvoiceNo) %>%

count(Country, sort = TRUE) %>%

slice\_head(n = 3) %>%

pull(Country)

Countries with most transaction counts:

1. United Kingdom 23494
2. Germany 603
3. France 461

Thereafter transaction list was creating using the StockCode and InvoiceNo from the top three countries.

Following code filters the dataset for a top country and keeps only unique combinations of InvoiceNo and StockCode to avoid duplicate items in the same transaction.

Thereafter it groups the items by invoice number to form a Apriori transaction list, which is converted into a transactions object required for Apriori analysis.

The Apriori algorithm is run with minimum support of 1%, confidence of 50%, and at least 2 items per rule. The resulting rules are sorted by lift to highlight the strongest associations.

for (country in top\_countries) {

cat("\n=====================\n")

cat("Running Apriori for:", country, "\n")

cat("=====================\n")

# Filter for country

country\_data <- data %>%

filter(Country == country) %>%

distinct(InvoiceNo, StockCode)

# Convert to transaction list and transactions object

trans\_list <- split(country\_data$StockCode, country\_data$InvoiceNo)

transactions <- as(trans\_list, "transactions")

# Run Apriori

rules <- apriori(transactions, parameter = list(supp = 0.01, conf = 0.5, minlen = 2))

rules <- sort(rules, by = "lift", decreasing = TRUE)

# Print readable rules

if (length(rules) == 0) {

cat("No rules found for", country, "\n")

} else {

# Convert rules to data frame

rule\_df <- as(rules, "data.frame")

rule\_df$LHS\_Desc <- get\_descriptions(lhs(rules), stock\_map)

rule\_df$RHS\_Desc <- get\_descriptions(rhs(rules), stock\_map)

cat("\nTop 10 rules by lift for", country, ":\n")

print(rule\_df %>%

select(LHS\_Desc, RHS\_Desc, support, confidence, lift) %>%

arrange(desc(lift)) %>%

head(20))

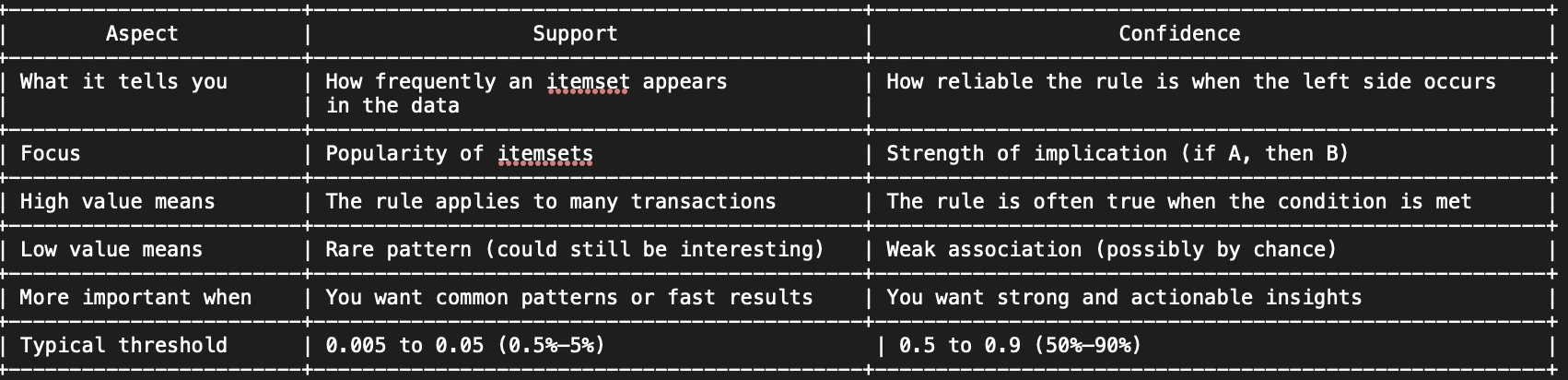
}

}

## Task 3 - Basket Analysis

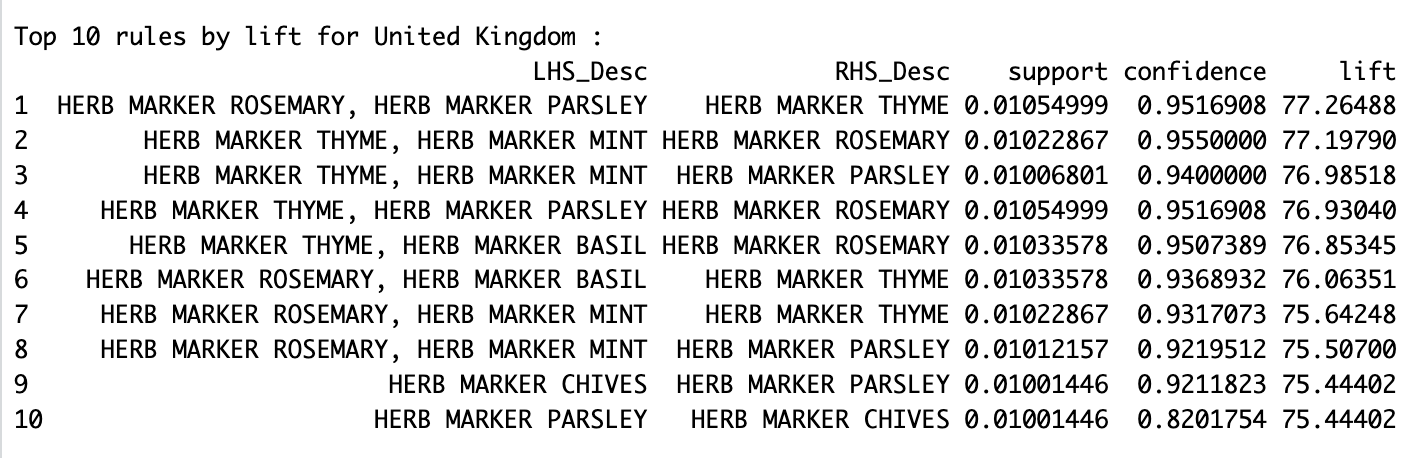
The initial Apriori run was done with a minimum support of 1%, confidence of 50%, and at least 2 items per rule. Although 70% - 80% will result strong reliable rules it will but it could lead to miss useful rules.

Support vs Confidence



Market Basket Analysis for United Kingdom

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2



Rule interpretations

1. It seems Customers who buy both ROSEMARY and PARSLEY almost always buy THYME too. Ideal for bundling.

2. If someone buys THYME and MINT, they are very likely also buy ROSEMARY. These three may form a strong trio of complementary items.

3. Similar to Rule 2 MINT and THYME pair strongly with PARSLEY. A different trio than Rule 2, but equally reliable.

4. Shows ROSEMARY is often bought when people buy PARSLEY and THYME. Reinforces prior rule.

5. Even substituting MINT/PARSLEY with BASIL still results in strong ROSEMARY prediction. ROSEMARY is a key crossover item.

6. ROSEMARY and BASIL jointly predict THYME. These three consistently show up across rules.

7. ROSEMARY + MINT again leads to THYME — confirms THYME’s strong linkage with most other herb markers.

8. MINT + ROSEMARY also predict PARSLEY with very high confidence.

9. CHIVES buyers very likely get PARSLEY — high lift suggests strong bond, despite single-item

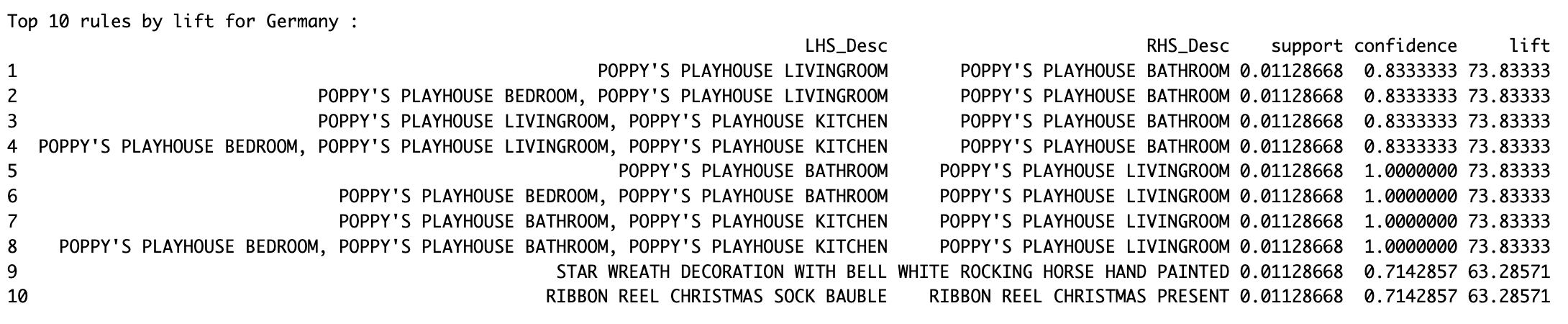
10. The reverse of Rule 9 — if someone buys PARSLEY, they frequently buy CHIVES too. Slightly lower confidence, same strong lift.

Overall, all these rules indicate extremely strong product associations within the “HERB MARKER” series of products. Almost all the rules’ projects confidence above 90% and lift above 75, which is exceptionally strong and incredibly high association strength. Seems like Herb markers are frequently purchased in combinations.

ROSEMARY, THYME, PARSLEY, and MINT consistently appear together, forming highly reliable combinations. Rules show that buying any two herbs almost guarantees purchase of a third. These findings are ideal for bundling and cross-selling.

Market Basket Analysis for Germany

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2 ordered for lift.



Overall, Germany shows strong associations within the "POPPY'S PLAYHOUSE" product line, especially between BATHROOM and LIVINGROOM sets.

Rules with 100% confidence suggest guaranteed purchases.

Holiday decorations like wreaths, baubles, and ribbon reels also exhibit strong bundling potential.

All 10 rules exhibit high lift values (63–74), indicating much higher co-occurrence.

Rule Interpretation

Rule 1 – 4

Customers who buy any playhouse room set almost always add the Bathroom set. Very strong cross-selling potential. With a confidence level of 83.33% and lift above 70% shows relatively strong and incredibly high association strength.

Rule 5–8:

With a confidence level of 100%, it seems Every customer who buys the BATHROOM set also buys the LIVINGROOM set. It's a guaranteed co-purchase highly valuable for promotions or recommendations.

Rule 9:

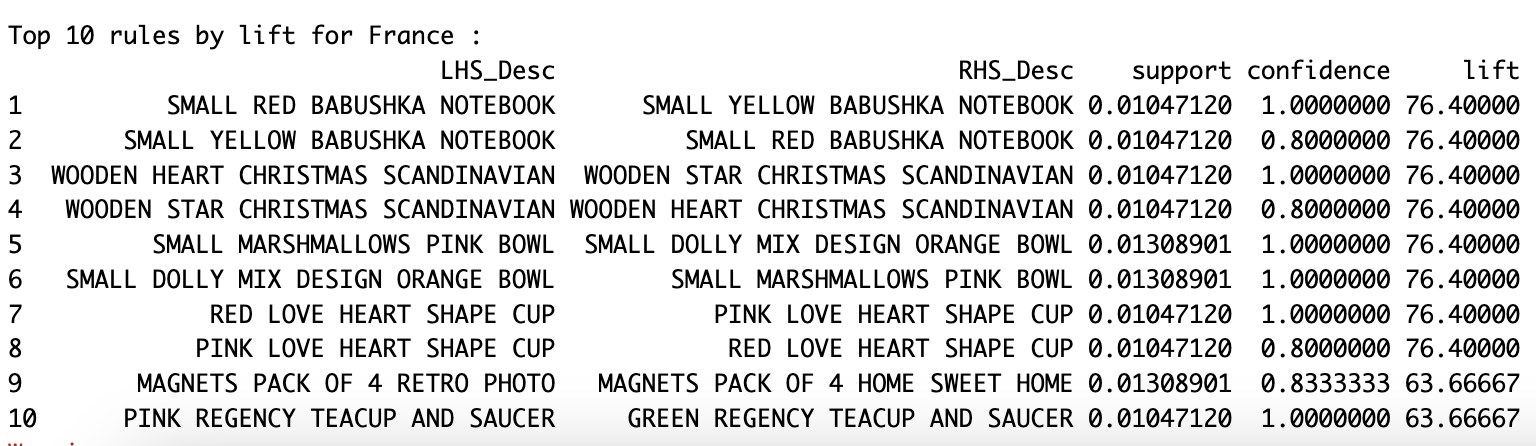
STAR WREATH DECORATION WITH BELL and WHITE ROCKING HORSE HAND PAINTED seems to be seasonal complementary items with strong association which are being bought together.

Rule 10

RIBBON REEL CHRISTMAS SOCK BAUBLE and RIBBON REEL CHRISTMAS PRESENT Christmas-themed ribbon items are highly associated. Strong opportunity for bundle deals during holiday seasons with a confidence level over 70%.

Market Basket Analysis for France

Apriori top 10 rules for parameters supp = 0.01, conf = 0.5, minlen = 2 ordered for lift.



Above rules depicts co-purchasing behavior among color or theme home variants. BABUSHKA notebooks, heart-shaped cups, and tea sets show 100% or near-perfect confidence, suggesting strong customer preference for matched or complementary products. These rules are ideal for bundling or gift-set curation.

Seasonal items like wooden Scandinavian decorations also pair reliably. Magnets and bowls reveal strong lift, reinforcing cross-sell value.

Rule Interpretation

Rule 1 – 2

SMALL RED BABUSHKA NOTEBOOK and SMALL YELLOW BABUSHKA NOTEBOOK seems to be highly correlated with a confidence value of 100% and lift of over 75%, therefore customers buying one BABUSHKA notebook very likely buy the other.

Rules 3 – 4

WOODEN HEART CHRISTMAS and WOODEN STAR CHRISTMAS highly correlated purchases with high confidence and lift values indicating frequent buys probably during the holidays. These two items can be seasonal and ideal to bundle together.

Rules 5 – 6

SMALL MARSHMALLOWS PINK BOWL and SMALL DOLLY MIX ORANGE BOWL seems to always be bought together which is ideal for bundling.

Rules 7 – 8

RED LOVE HEART CUP and PINK LOVE HEART CUP also being bought together just like above items. Mostly like as a pair or a set.

Rule 9

MAGNETS PACK OF 4 RETRO PHOTO likely to by HOME SWEET HOME product item.

Rule 10

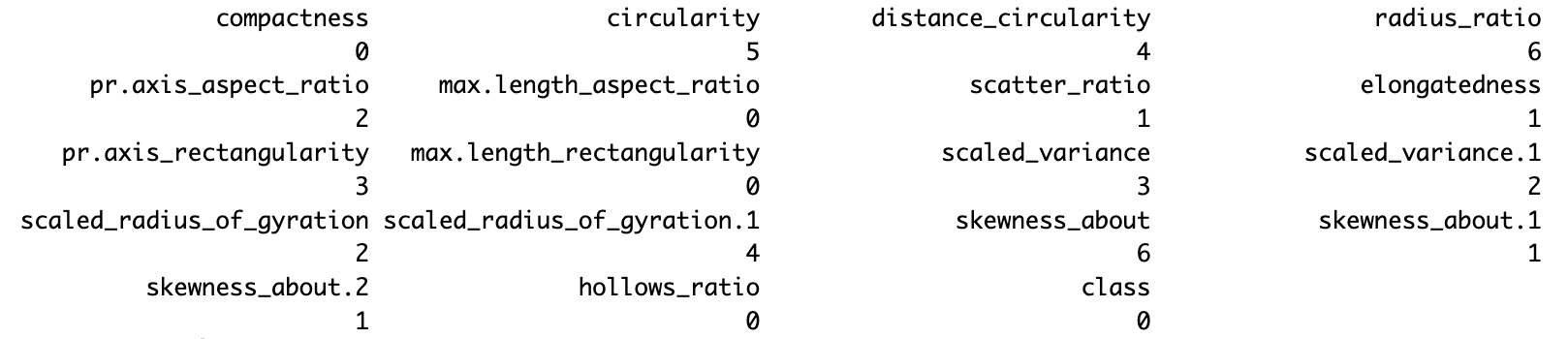
PINK REGENCY TEACUP AND SAUCER and GREEN REGENCY TEACUP AND SAUCER a guaranteed co-purchase with confidence level of 100%. Seems like customers buying the pink variant always buy the green one, suggesting strong interest in color-matched tea sets.

Question 2: Clustering

## Task 1. K-Means clustering after performing dimensionality reduction (PCA)

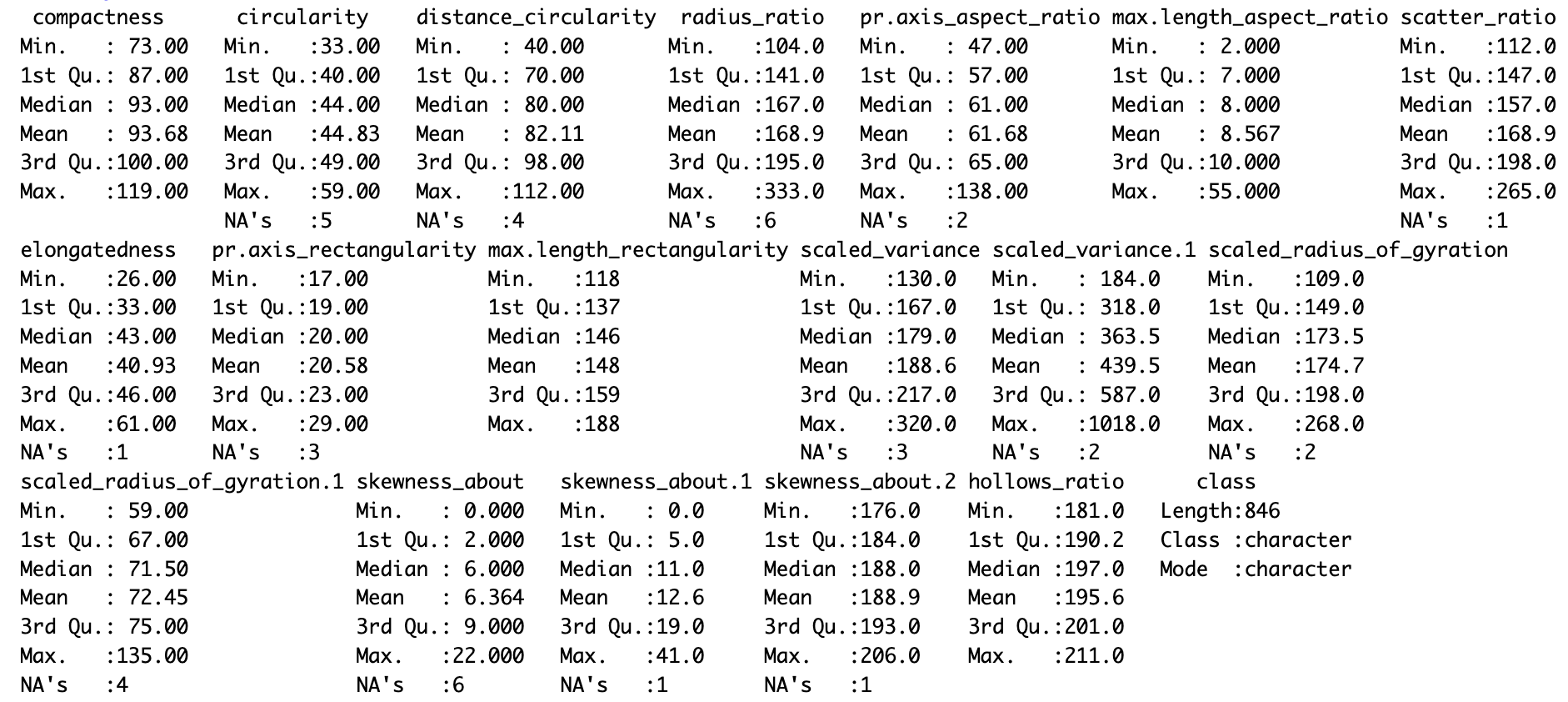
Following things were done to achieve to K Means clustering.

1. First data set is loaded to RStudio.
2. Verified that dataset contains all the columns, data types and then analyzed for missing values.



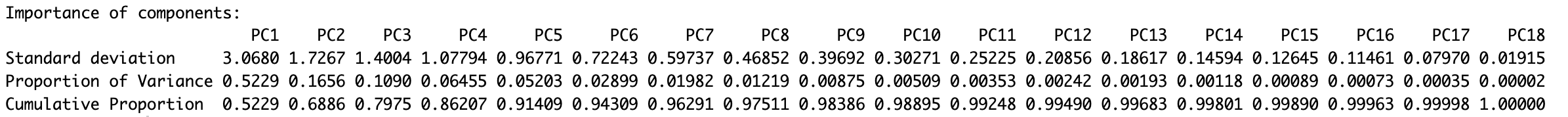
Each column had few missing values, and it was decided to remove those. This removed 33 entries from the dataset.

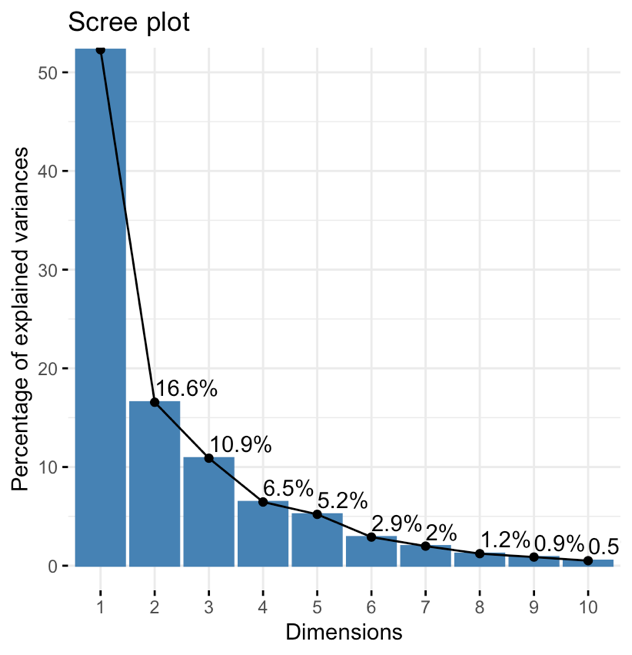
1. Then dataset was analyzed to find outliers, min, max and mean values.



It seems there aren’t any extreme outliers, however it was decided to remove outliers using IQR method. This will be beneficial as PCA is based on variance and covariance, and having outliers may orient principal components toward those outliers which could lead to misleading directions and poor clustering. Therefore, having outliers skewed toward a few distant points.

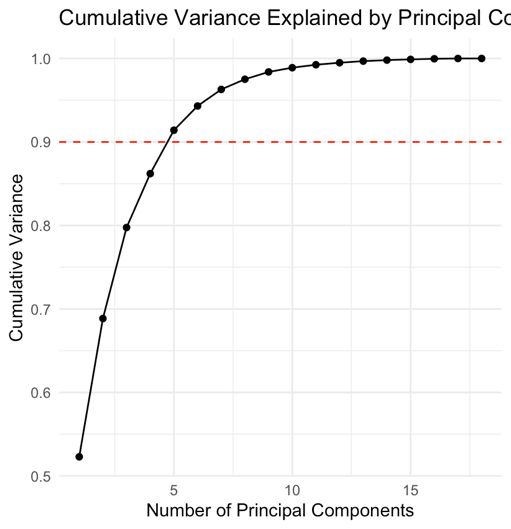
1. After outlier removal feature scaling was done for standardization of columns. So PCA won't be biased to larger values and K-Means won’t treat large-scale features as more prominent.
2. Then dimensionality reduction was done using the prcomp() method.



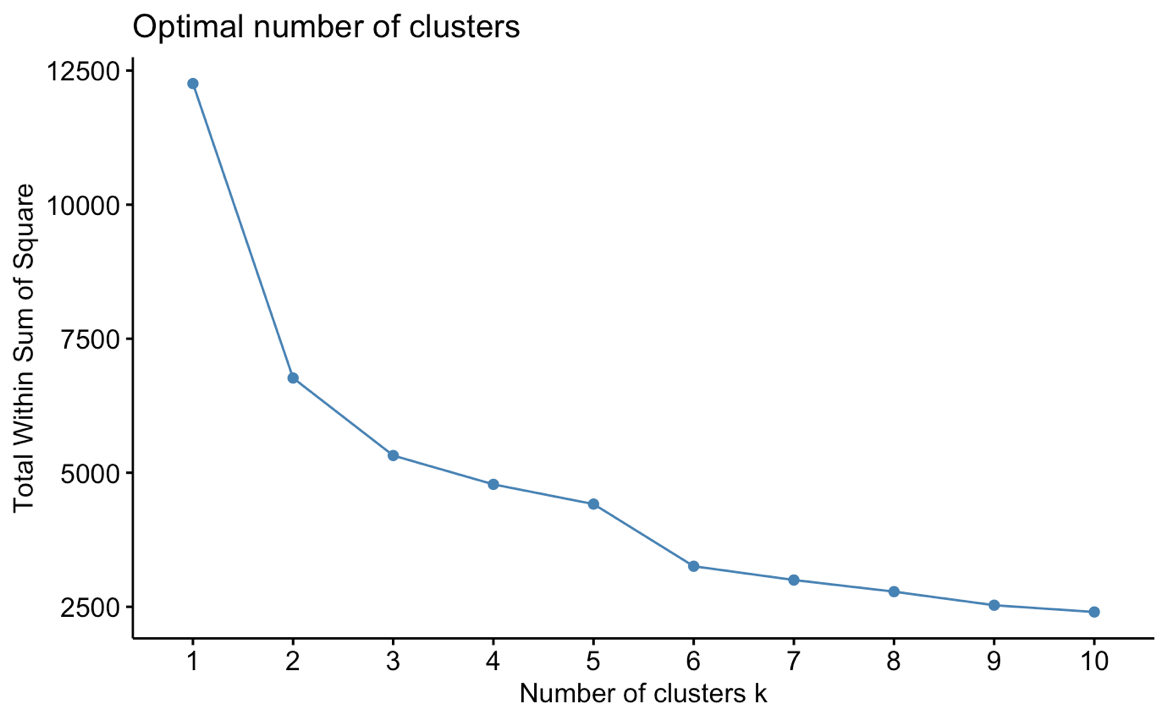


## According to the PCA results, PC1 alone captures over 52% of the variance, while PC1 through PC5 collectively account for more than 91% of the total variance in the dataset. This indicates that most of the structural information in the vehicle silhouette features can be effectively represented in a reduced 5-dimensional space. Therefore, dimensionality reduction to five principal components prior to clustering is both justified and effective, as it retains most of the original variance while simplifying the feature space.

1. Also, this further proves by observing the cumulative variance plot. Below cumulative variance plot, the black curve represents the cumulative variance, and the red dashed line represents the 90% threshold. The black curve crosses the red line between the 4th and 5th components indicating that the first five principal components capture over 90% of the total variance in the dataset.



1. Then number of optimal clusters required for the K-means was found out that using the elbow method.



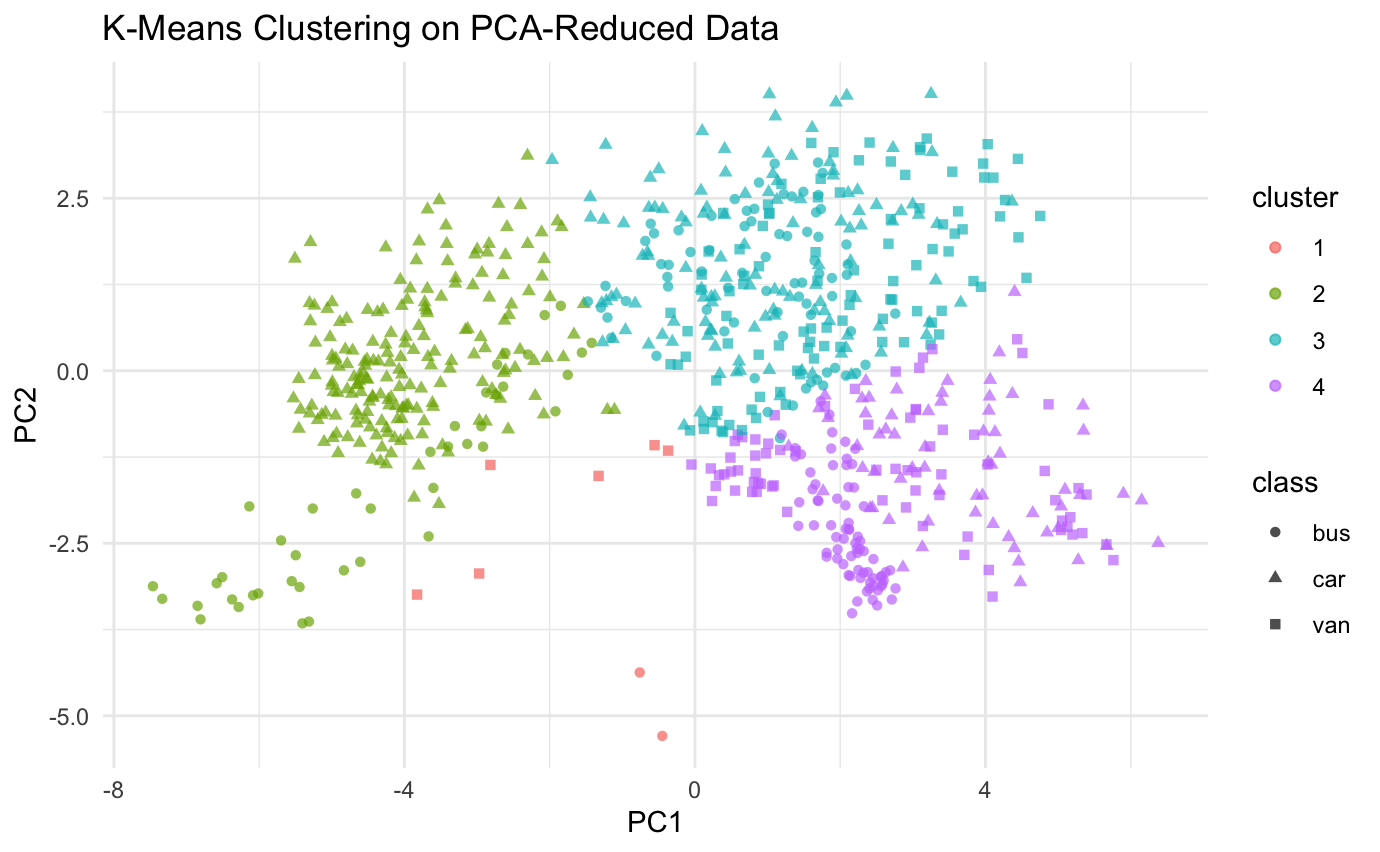
According to the plot the most significant drop occurs between k = 1 and k = 4.

After k = 4, drop slows down. Between K= 4 to k = 6 drop is still there, but at a slower pace.

Therefore, the most noticeable "elbow" occurs around k = 4, K = 6 also acceptable if granular details are required.

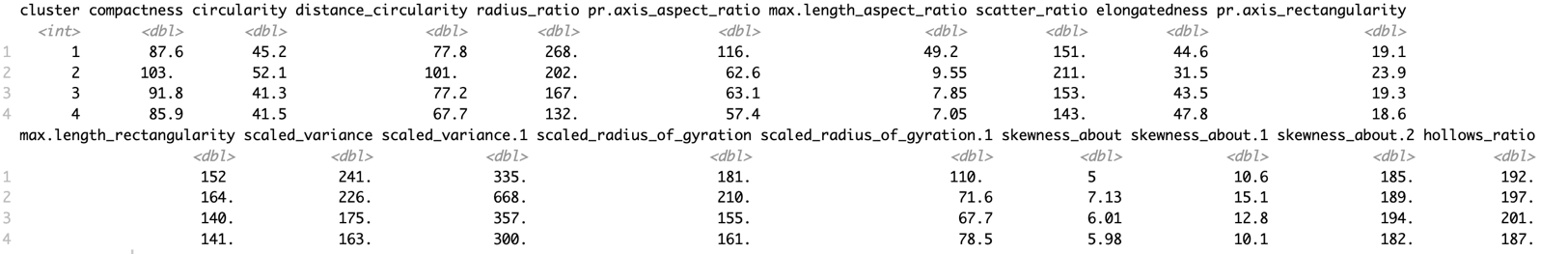
So optimal clusters can be 4 or 6.

1. When 4 clusters were used.



PC1 and PC2 were chosen for plotting because they capture the most variance in the data (~69%).

Cluster Profile



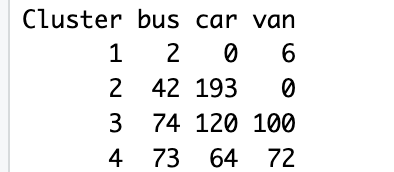
Cluster profiling table summarizes the average values of key shape features for each cluster after K-Means clustering (k = 4).

Cluster 2 (Car-dominant): Highest compactness (103) and circularity (52.1) also indicates symmetrical and compact silhouettes.

Cluster 3 (Mixed): Moderate on all matrix indicating silhouettes with irregular or less symmetric shapes.

Cluster 4 (Bus/Van-heavy): Lowest compactness (85.9) and radius ratio (132), Suggests boxier, more rectangular shapes

Confusion Matrix



### Conclusion

The vehicle dataset included properties such as scaled variance, skewness, kurtosis, compactness, circularity, rectangularity etc, and the vehicles include double-decker bus, Chevrolet van, and two car models: Saab and Opel Manta.

The experimental design anticipated that the bus, van, and one of the car types would be distinguishable, whereas the two car models might be more difficult to separate.

**Cluster 1 (Outlier / red cluster):**

This is a sparse group which contains a small number of vehicles from various classes. These likely represent atypical silhouettes or rare viewing angles, and they do not align strongly with any single vehicle type.

**Cluster 2 (Car-dominant / green):**

A condensed, homogeneous cluster composed almost entirely of cars.

The Cluster 2 lies in the low-PC1 region, and PC1 has strong negative loadings for compactness (−0.274), circularity (−0.294), as well as distance\_circularity (−0.305) and scaled\_variance (−0.310). These properties suggest that vehicles in this cluster have tight, symmetric silhouettes with low shape variability, consistent with compact car profiles.

**Cluster 3 (Mixed core cluster / cyan):**

The most compositionally diverse cluster, containing a substantial mix of cars, vans, and buses. This cluster represents vehicles with overlapping shape profiles and supports the hypothesis that distinguishing between the two car types is challenging.

The Cluster 3 is centrally located in the PCA plot and contains a balanced mix of cars, buses, and vans. Its position near the origin suggests moderate PC1 scores, indicating average values across features such as compactness, circularity, distance\_circularity, and scaled\_variance. Since PC1 reflects these features with negative loadings, vehicles in this cluster likely have neither strongly compact nor highly elongated silhouettes, but instead intermediate, structurally overlapping shapes. This supports the observed class mixing and aligns with the assignment expectation that distinguishing between certain car types would be challenging.

**Cluster 4 (Bus and van cluster / purple):**

Cluster 4 appears in the higher-PC1 region of the PCA plot. Since PC1 is negatively weighted on features such as compactness (−0.274), circularity (−0.294), distance\_circularity (−0.305), and scaled\_variance (−0.310), a high PC1 score suggests lower values for these features. This indicates that vehicles in this cluster tend to have elongated or rectangular silhouettes, which aligns well with bus and van profiles.

However, the presence of cars within this cluster suggests that some car silhouettes overlap structurally with vans or buses — likely due to angle of view or shared geometric traits (e.g., side views with lower circularity). Thus, Cluster 4 may represent larger or less compact silhouettes across multiple classes, rather than exclusively capturing specific vehicle types.

## Task 2 - Agglomerative hierarchical clustering after performing dimensionality reduction (PCA)

Code : <https://github.com/lolitha-lakshan-1/DataMining_assignment_1/blob/main/Clustering.Rmd>

Code to perform Agglomerative hierarchical clustering

# Compute distance

# Euclidean distance reflects how far two points are from each other in terms of overall variance-base and data is continous

distance\_matrix <- dist(pca\_selected, method = "euclidean")

#Agglomerative Clustering

# Ward method works only with squared Euclidean distances

hc <- hclust(distance\_matrix, method = "ward.D2")

#Plot the Dendrogram

plot(hc, labels = FALSE, hang = -1, main = "Hierarchical Clustering Dendrogram")

#Cut Dendrogram into k Clusters

cluster\_cut <- cutree(hc, k = 4)

# Plot PCA

pca\_df <- as.data.frame(pca\_selected)

pca\_df$cluster <- factor(cluster\_cut)

pca\_df$class <- vehicles\_no\_outliers$class # Optional: for coloring by true class

library(ggplot2)

ggplot(pca\_df, aes(x = PC1, y = PC2, color = cluster, shape = class)) +

geom\_point(alpha = 0.7) +

labs(title = "Hierarchical Clustering (Ward) on PCA-Reduced Data",

x = "PC1", y = "PC2") +

theme\_minimal()

# Confusion Matrix

table(Cluster = pca\_df$cluster, Class = pca\_df$class)

Above code applies after reducing the dataset using PCA to five principal components and then euclidean distance matrix is computed to measure dissimilarity between observations.

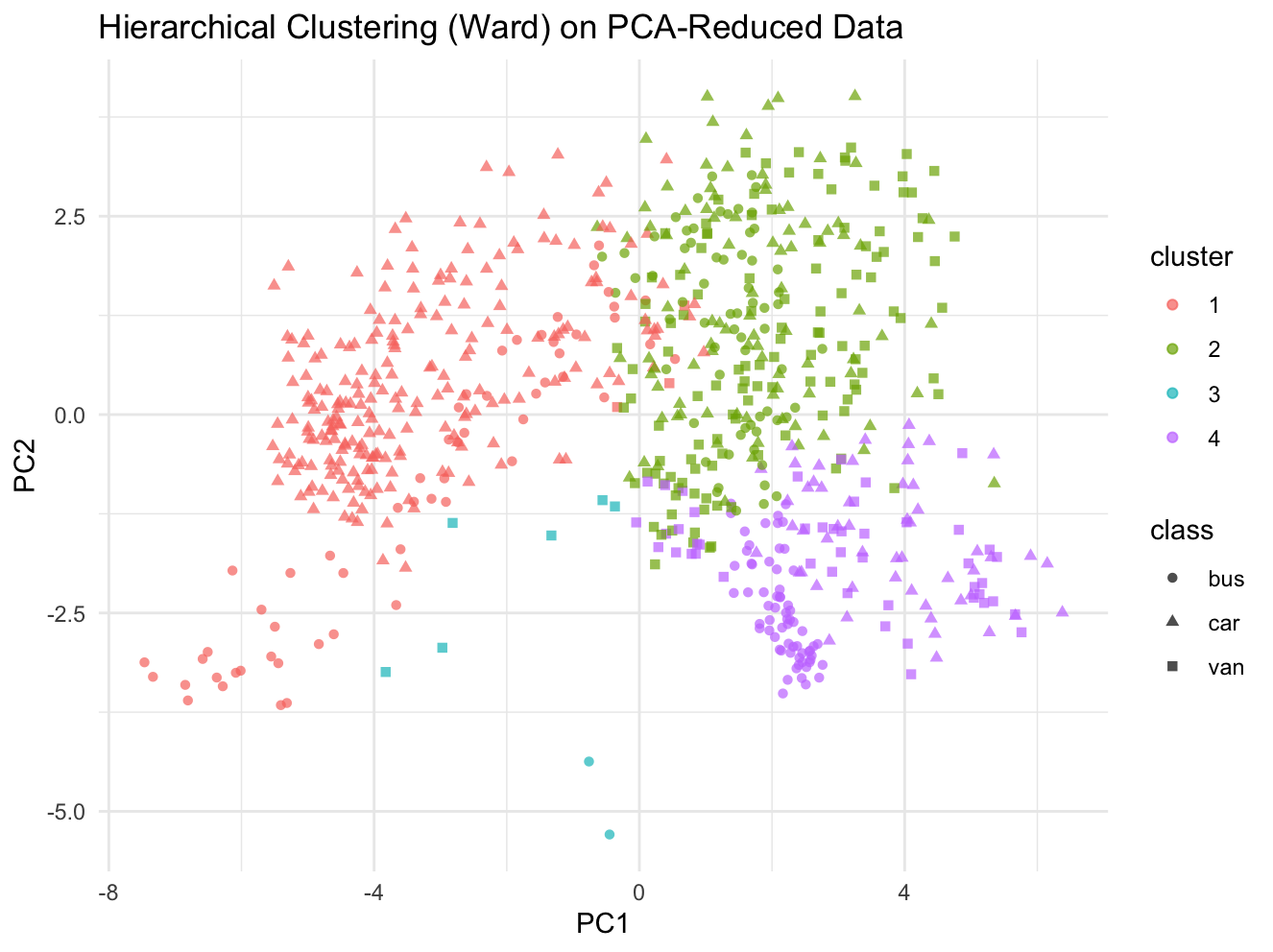
Thereafter, Ward.D2 method is used to perform agglomerative clustering. A dendrogram is plotted to visualize how clusters merge at different distance levels. The tree is then cut at a chosen height to form four clusters.

A diagram of a clustering structure

AI-generated content may be incorrect.

Above dendrogram depicts hierarchical structure in the PCA-reduced vehicle data. The biggest jump in height occurs around height 40–50, indicating 4 major clusters. This also proves the decision to use 4 clusters in K-Means as well.

The tree also shows that some vehicle shapes are grouped closely together in small subgroups, especially near the bottom of the plot.



Confusion Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **Bus** | **Car** | **Van** | **Dominant Class** | **Explanation** |
| 1 | 57 | 231 | 3 | Cars | Clear separation of cars, preferably one car model, Saab or Opel. High compactness and circularity implied. |
| 2 | 66 | 93 | 123 | Highly Mixed | buses (66), cars (93), and vans (123), supports hypothesis that distinguishing between car types and van is difficult. Represents overlapping silhouettes. |
| 3 | 2 | 0 | 6 | Van | Outliers, edge cases or rare shapes. |
| 4 | 66 | 53 | 46 | Mixed | Possibly boxier, rectangular silhouettes because positioned on far right of PCA plot (high PC1). |

### Conclusion

## Cluster 1 successfully isolated most cars (231 out of 327) and is in the low-PC1 region of the plot. Since PC1 has strong negative loadings for compactness (−0.274) and circularity (−0.294), low PC1 scores correspond to higher values for these features in the original data. This supports the claim that vehicles in Cluster 1 exhibit high compactness and symmetry, typical of compact car silhouettes.

## Cluster 4 is positioned on the high end of PC1, and due to the same negative loadings, high PC1 values indicate low compactness and circularity. This confirms that Cluster 4 contains vehicles with more elongated or rectangular silhouettes, which aligns with the presence of many buses and vans.

## Cluster 2 showed substantial class mixing, suggesting overlapping shape profiles between vehicles.

## Cluster 3 contained only a few observations and appears to represent rare or outlier shapes.

## Overall, the clustering results support the hypothesis that buses and vans are distinguishable based on shape, while separating car models remains more challenging.

## Task 3

The cluster analysis helps to identify vehicle categories such as buses, vans, and cars based on shape. This is important for camera-based classification systems, autonomous driving, smart traffic monitoring, and automated tolling where quick and reliable silhouette recognition is crucial.

The analysis also highlights outlier or rare shapes, enabling systems to flag them for special handling.

It also validates the assumption that buses and vans are more distinguishable than similar-looking cars, which can guide model training. Overall, this enhances the efficiency, scalability, and real-world usefulness of shape-based vehicle detection solutions.